NPS55-85-012

NAVAL POSTGRADUATE SCHOOL

Monterey, California

AN EXAMINATION OF SOME ERROR CORRECTING

TECHNIQUES FOR CONTINOUS SPEECH RECOGNITION

TECHNOLOGY

by

Gary K Poock
B. Jay Martin

June 1985

Approved for public release; distribution unlimited.

Prepared for:
Naval Ocean Systems Center
Code 81
San Diego, CA 92152

FedDocs
D 208.14/2
NPS-55-85-012

I 402.12:NPS-55-85-012

NAVAL POSTGRADUATE SCHOOL
Monterey, California

Rear Admiral R. H. Shumaker
Superintendent

David A. Schrady
Provost

This research was supported and funded by the Naval Ocean Systems Center.

DUDLEY KNOX LIBRARY
NAVAL POSTGRADUATE SCHOOL
MONTEREY CA 93943-5101

Unclassified
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: NPS55-85-012
2. GOVT ACCESSION NO.:
3. RECIPIENT'S CATALOG NUMBER:

4. TITLE (and Subtitle): AN EXAMINATION OF SOME ERROR CORRECTING TECHNIQUES FOR CONTINUOUS SPEECH RECOGNITION TECHNOLOGY

5. TYPE OF REPORT & PERIOD COVERED: Technical

6. PERFORMING ORG. REPORT NUMBER:

7. AUTHOR(s): Gary K. Poock
B. Jay Martin

8. CONTRACT OR GRANT NUMBER(s): N62271-84-M-3326

9. PERFORMING ORGANIZATION NAME AND ADDRESS: Naval Postgraduate School
Monterey, CA 93943-5100

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS:

11. CONTROLLING OFFICE NAME AND ADDRESS: Naval Ocean Systems Center
Attn: Code 421
San Diego, CA 92152

12. REPORT DATE: June 1985

13. NUMBER OF PAGES: 37 pages

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office):

15. SECURITY CLASS. (of this report): Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE:

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

VTAG, Voice Recognition, Speech Recognition, Error Correction

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Continuous automatic speech recognition systems are becoming more available and obtaining fairly good recognition accuracies of connected speech input. The question now arises as to how one corrects errors. For example, if one entered 20 digits and all were correct but the 12th one, how would you correct that error? This research examined 5 different techniques for using automatic continuous speech input to correct previous errors made by the automatic speech recognition system.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S N 0102-LF-014-6601

Unclassified
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

CONTENTS

EXECUTIVE SUMMARY

The primary purpose of this research was to examine the effects of various input string lengths and error correction methods on the recognition accuracy and efficiency of a currently available continuous automatic speech recognition (ASR) system. The effect of sex was examined also and an estimate of the average recognition accuracy of a continuous ASR system was sought.

In the entry of numerical data, the input string length of seven digits at a time proved significantly more efficient than strings of three or five. Although subjects preferred some error correction methods over others, there were no significant differences in error rates or efficiency due to the correction method used. There were also no significant differences due to sex.

The average recognition accuracy of the continuous ASR system was conservatively estimated at over 95%. These findings and areas of possible future research are discussed.

1. INTRODUCTION

1.1 Background

In recent years, voice technology has developed to the extent that basic systems have now been used successfully in several industrial and military applications. Voice recognition devices that have been installed in "real world" situations have reduced input errors, cut task time, increased user friendliness, and proven cost effective in general (Nye, 1982; Poock, 1982). This successful climate, along with continued reductions in the cost of voice recognition systems, has made voice input an attractive alternative to motor input in a wide variety of settings.

Until recently, reliable ASR has been confined to recognition of discrete speech, that is, utterances of up to about two seconds in length and pauses of about 150 ms. between utterances. With the advent of continuous ASR systems the interactive process of ASR may be faster and more natural, increasing the efficiency of ASR and its potential applications. However, as with any new technology, new questions and issues need to be addressed. The most basic issue in continuous ASR (and ASR in general) is system efficiency. In effect, what input speed and accuracy can be expected in the operation of a continuous ASR system? As is the case with discrete ASR, the answer to this question can vary widely from one application to another. In particular, the type of vocabulary can significantly affect recognition accuracy (Armstrong and Poock, 1981). The vocabulary consisting of digits, zero through nine, warrants special attention due to its frequency of use across applications. Therefore, a study concerning numerical data entry via continuous ASR should prove most useful in terms of measuring the baseline recognition accuracy of the system and in generalization of the results to other applications.

1.2 Problem

In the context of discrete ASR an investigation of numerical data entry would be fairly straight forward. For example, a discrete digit is spoken and the ASR system displays feedback of the match or mismatch. In the case of a mismatch the speaker would immediately cancel the error with some key word like "erase" or "rubout," and then try again. System efficiency would be measured in terms of average input speed and accuracy.

With the capabilities of continuous ASR the investigation becomes somewhat more complex. The first issue concerns the number of digits to constitute an input -- since a truly continuous ASR system could accept any number of digits, from one to infinity, as a single input for which it is to produce a matching set. Assuming a fixed number of total digis (e.g., 50) will be input, different individual input string lengths may result in different speed and accuracy rates. The input of 25 two-digit strings would reqire 24 inter-string pauses for recognition and feedback, compared to only four such pauses with the input of five ten-digit strings.

Coarticulation may also be a factor in string length. Coarticulation is the simultaneous pronunciation of the end of one word and the beginning of another, e.g., "three-eight." The input of a two-digit string requires the ASR system to deal with only one coarticulation in finding the boundary between the two words. However, a ten-digit string requires the processing of 9 coarticulations.

Without a specific application in mind the determination of input string lengths for investigation becomes somewhat arbitrary. However, a speaker's short term memory for digits should be about seven, give or take a couple

(Miller, 1956), placing a limit on the number of digits he or she can comfortably remember for input and mentally compare to output. Based on this assumption and practial considerations, the input string lengths of three, five, and seven, were chosen for investigation.

Another issue is error correction. In the discrete ASR of digits each output is either 100% right or 100% wrong. However, with continuous ASR an output may be partially incorrect. For example, the input string is "1, 4, 3, 5, 2" and the output is "1, 4, 3, 9, 2." These errors could be handled like discrete ASR errors, in which case the speaker would isue the "ERASE" or "RUBOUT" command and try again. However, other methods of correction that address only the incorrect portion of the output may be more efficient. In the example above it may be faster to change the 9 to a 5 than to erase the entire output and repeat the whole string again. In addition, addressing only the specific error (changing the 9 to a 5) gives the ASR sysem a different speech input to correct the error (e.g., "CHANGE THE 9 TO A FIVE") rather than the same speech input ("1, 4, 3, 5, 2") which has already demonstrated a propensity for misrecognition.

The question then, is what are the alternative correction methods for partial errors? The possibilities are limited only by one's imagination and degree of control over the feedback display. Four error correction methods were chosen for use in the experiment:

1) "RUBOUT" - erases the entire output regardless of partial or total error.

E.g., Input = 1, 4, 3, 5, 2
Output = 1, 4, 3, 9, 2 Subject says "RUBOUT," "1, 4, 3, 5, 2"

2) "POSITION X MAKE-IT Y" - changes xth digit (from left to right) to Y

E.g., Input = 1, 4, 3, 5, 2
Output = 1, 4, 3, 9, 2 Subject says "POSITION 4 MAKE-IT 5"

3) "BACKUP X MAKE-IT Y" - changes xth digit (from right to left) to Y

E.g., Input = 1, 4, 3, 5, 2
Output = 1, 4, 3, 9, 2 Subject says "BACKUP 2 MAKE-IT 5"

4) "CHANGE X (nth ONE) MAKE-IT Y" - changes the nth X to Y, if n is not stated then the first X (from left to right) is changed to Y.

E.g., Input = 1, 4, 3, 5, 2
Output = 1, 4, 3, 9, 2 Subject says "CHANGE 9 MAKE-IT 5"

E.g., Input = 1, 9, 3, 5, 2
Output = 1, 9, 3, 9, 2 Subject says "CHANGE 9 SECOND-ONE MAKE-IT 5"

NOTE: Subjects use contiruous speech to say everything in the examples above.

1.3 Objectives

The specific objectives of this research were as follows:

(1) To examine the effects of 3 different input string lengths on continuous ASR accuracy and efficiency.

(2) To examine the effects of four different correction methods on continuous ASR efficiency.

(3) To examine any interaction effects of the three string lengths with the four correction methods in terms of accuracy and efficiency.

(4) To obtain an estimate of the recognition accuracy of a currently available continuous ASR device.

(5) To examine the effects, if any, of gender on accuracy and efficiency.

2. METHOD

2.1 Subjects

Twelve volunteers were recruited primarily from the Naval Postgraduate School in Monterey, CA. Six males included 4 Naval officers, 1 Marine officer, and 1 National Reservist, 4 secretaries, and 1 elementary school teacher not associated with the Naval Postgraduate School. One subject had worked with ASR for about 3 years. Three subjects had about 3 hours of experience each with a discrete ASR system and the remaining 8 subjects had never used an ASR system. Five subjects had previous microphone experience as pilots, navigators, or radio operators.

2.2 Apparatus

A Verbex 3000 continuous ASR system was used in this study. The system is capable of recognizig natural continuous speech of indefinite length, limited only by an output buffer of 240 characters per recognition set.

A Shure model SM12A headset microphone was used as the input device. This microphone is supplied as standard equipment with the Verbex.

Prompts and recognition sets were displayed on Lear Siegler ADM31 video display terminal.

2.3 Experimental Design

This experiment employed a 5 x 3 x 2 mixed design. Five correction methods were crossed with three input string lengths. The correction methods were RUBOUT, POSITION, BACKUP, CHANGE, and ALL -- in which the previous four methods were all available. The input string lengths were 3, 5, and 7

digits. Two groups of subjects, 6 males and 6 females, constituted the between subjects variable, and experienced all combinations of correction methods and input string length. A summary of the experimental design appears in Figure 2-1.

2.4 Procedure

2.4.1 Introduction. The experiment was divided into a training session lasting 45 minutes and a test session of 45 minutes. Subjects signed up for the individual 45 minute sessions at their convenience. Seven subjects did the training and testing sessions on separate days no more than one week apart.

The sessions took place in the C^3 lab at the Naval Postgraduate School. The Verbex was located in a 18 by 16 foot acoustically paneled room with several other computer terminals and peripherals. During the course of any session it was common to have several people talking and typing in the room. Also located in the room was a heavy (lead shielded) pneumatically operating sliding door. The opening and closing of this door produced significant noise in the room, peeking at approximately 85 dbC, and as the main entrance to the lab, it was opened and closed frequently. Although the various sources of noise were considerable, no measures were taken to reduce or control them. The resulting sound level environment ranged from 64 to 85 dbC with a mode of about 72 dbC.

At the beginning of each subjects' first session the experimenter described the experiment and gave a demonstration of how the continuous ASR system would later be used by the subject for numerical data entry and error correction.

2.4.2 Training. After the demonstration the experimenter led the subject through the training phase. The term "training" as used in ASR, refers to

		CORRECTION METHOD														
		RUBOUT			POSITION			BACKUP			CHANGE			ALL		
GENDER	INPUT STRING LENGTH / SUBJECT #	3	5	7	3	5	7	3	5	7	3	5	7	3	5	7
MALE	1															
	2															
	3															
	4															
	5															
	6															
FEMALE	7															
	8															
	9															
	10															
	11															
	12															

FIGURE 2-1.
SUMMARY OF EXPERIMENTAL DESIGN

the process by which the speaker makes known to the recognizer the characteristics of his/her particular speech patterns for all the utterances he/she will be using. Twenty utterances were used in the current study (see Appendix A). For the Verbex 3000 this training procedure consists of two phases, isolated and continuous. In the isolated training phase the speaker says each utterance in the vocabulary at least twice by itself (discrete or isolated). Isolated training was suspended when the pneumatic door was activated since Verbex recommends isolated training in a quiet environment.

In the continuous training phase up to three utterances are grouped together and spoken continuously. Each utterance was included in about 20 such groups and was therefore coarticulated about 20 times during the continuous training phase. Two hundred such groups of utterances were spoken. Subjects were reminded to speak in a natural voice, but somewhat more quickly than in normal conversation, since they would be speaking rapidly in the subsequent test session. Continuous training proceeded throughout noises from the pneumatic door and talking in the room. The continuous training phase took approximately 25 minutes per subject.

As a result of these training phases the Verbex retains a template in memory on each utterance. Ideally, subsequent utterances (in testing) are matched with the template for the same utterance in memory, resulting in a correct recognition and output. In cases where a match is not found, a nonrecognition or rejection occurs and the Verbex makes no response. Occasionally, the recognizer makes an incorrect match and an incorrect response is output, constituting a misrecognition or misinterpretation of the utterance.

2.4.3 Testing. Before data collection each subject completed a practice session. In the practice session a randomly generated five digit prompt appeared in the upper right-hnd corner of the display screen. The subject spoke the digits and the recognition set was output directly below the prompt. If the recognition was 100% correct, the screen cleared after 2.9 seconds and the process was repeated with a new prompt. If any part of the output was incorrect the display remained the same until a correction command was entered. The output string was then immediately modified to reflect the correction. If no further corrections were necessary the screen cleared and a new prompt appeared in 2.9 sec. All four correction methods were available. The subject was to say "RESTART" whenever the Verbex produced no output to the subjects input. An audible beep signaled recognition of the "RESTART" command.

Subjects were reminded that in the test phase total input time would be measured and were instructed to test the limits of te ASR system during practice by speeding up their inputs until they resulted in output errors. This gave the subjects a good estimate of how fast they could enter the digits as well as the cost (in time) of correcting errors. Each of the correction methods was practiced until the subjects demonstrated a clear understanding of, and ability to quickly execute, all of them.

Once the subject completed practice and the experimenter answered any questions the data collection began. The task was to input a total of 105 digits, with corrections, in as short a time as possible. Each subject entered 105 digits -- 3 at a time, 5 at a time, and 7 at a time -- under each of the 5 correction conditions. The order of the four correction methods -- RUBOUT, POSITION, BACKUP, and CHANGE were counter balanced for both sequential position and preceding condition (Bradley, 1976). The ALL correction condition was always done last. For each subject, five of the six possible string length sequences were chosen randomly and randomly ordered across the five correction conditions.

In testing, the prompt of 3, 5, or 7 digits appeared in the upper right hand corner of the screen. The subject then said the digits. If the system could not find a match or only "heard" a portion of the input, it made no response, in which case the subject would say "RESTART," hear a beep, and try again. If there was an error in the recognition output, the output string was displayed directly below the prompt and remained there until a correction command was recognized. The output string was immediately modified to reflect the correction. If "RUBOUT" was used the output string was replaced by dashes (- - -). Once the output string matched the prompt (whether on the original input or after one or more corrections) the screen was immediately cleared of both prompt and output and a new prompt appeared. The process of recognition, accuracy checking, screen clearing and presenting a new prompt took 0.8 seconds in all conditions. This process was repeated until a total of 105 digits had been correctly recognized. The experimenter timed each run and the computer (in the Verbex) tracked and reported the number of errors (corrections and nonrecognitions). The experimenter recorded the time and errors at the end of each run.

After the subjects completed all the conditions they were asked if they preferred any correction method(s) over the others, and if there was any correction method they thought was either so useless or confusing that they would not even make it an option. Responses to these questions were recorded by the experimenter and the test phase was concluded.

2.5 Independent and Dependent Variables

The independent variables in this study were input string length (3, 5, 7); error correction method (RUBOUT, POSITION, BACKUP, CHANGE, all); and sex. The dependent variables were efficiency (time to input 105 digits correctly) and recognition accuracy.

3. RESULTS

3.1 Overview

For error data all analyses of variance procedures and post hoc range tests were performed using the arcsin transformation of raw data to stabilize the variance of the error terms (Neter and Wasserman, 1974). The mean error and time rates that appear in the tables and figures are untransformed. All a posteriori tests for significance between pairs of means were performed using the Scheffe procedures described in Bruning and Kintz (1977).

Section 3.2 presents the data on efficiency (time to correctly recognize 105 digits). Section 3.3 presents data on total errors. Section 3.4 presents data on subjects responses to post test questions.

3.2 Efficiency

Table 3-1 presents the analysis of variance for efficiency (time to correctly recognize 105 digits). A significant main effect of input string length was found ($F = 25.059$, $p < .001$). No other main effects or interactions were significant. Mean total time (in seconds) for input string length by correction method are shown in Table 3-2. The main effect of input string length is shown graphically in Figure 3-1.

Scheffe tests were performed to detect single effects between input string lengths. Inputting digits seven at a time was significantly more efficient than both five at a time and three at a time, at the $p < .05$ level. Inputting 5 digits at a time was significantly more effiient than 3 at a time at $p < .1$ level.

TABLE 3-1

ANALYSIS OF VARIANCE SUMMARY TABLE
OF EFFICIENCY

SOURCE	df	MS	F
GENDER (G)	1	2191.022	.442
ERROR	10	4958.616	
CORRECTION METHOD (C)	4	635.506	.874
C G	4	764.106	1.051
ERROR	40	727.249	
INPUT STRING LENGTH (L)	2	11242.839	25.059 *
L G	2	134.372	.300
ERROR	20	448.652	
L C	8	436.131	.899
L C G	8	304.164	.627
ERROR	80	485.186	

* P<.001

TABLE 3-2

MEAN TOTAL TIME (IN SECONDS) FOR
INPUT STRING LENGTH BY CORRECTION METHOD

		INPUT STRING LENGHT			
		3	5	7	$\overline{X}$ CORRECTION METHOD
CORRECTION METHOD	RUBOUT	107.67	88.50	73.08	89.75
	POSITION	98.42	80.58	76.50	85.17
	BACK-UP	102.08	105.67	82.00	96.58
	CHANGE	105.50	100.92	72.50	92.97
	ALL	102.83	93.42	77.50	91.25
	$\overline{X}$ INPUT STRING LENGTH	103.30	93.82	76.32	91.14 GRAND $\overline{X}$

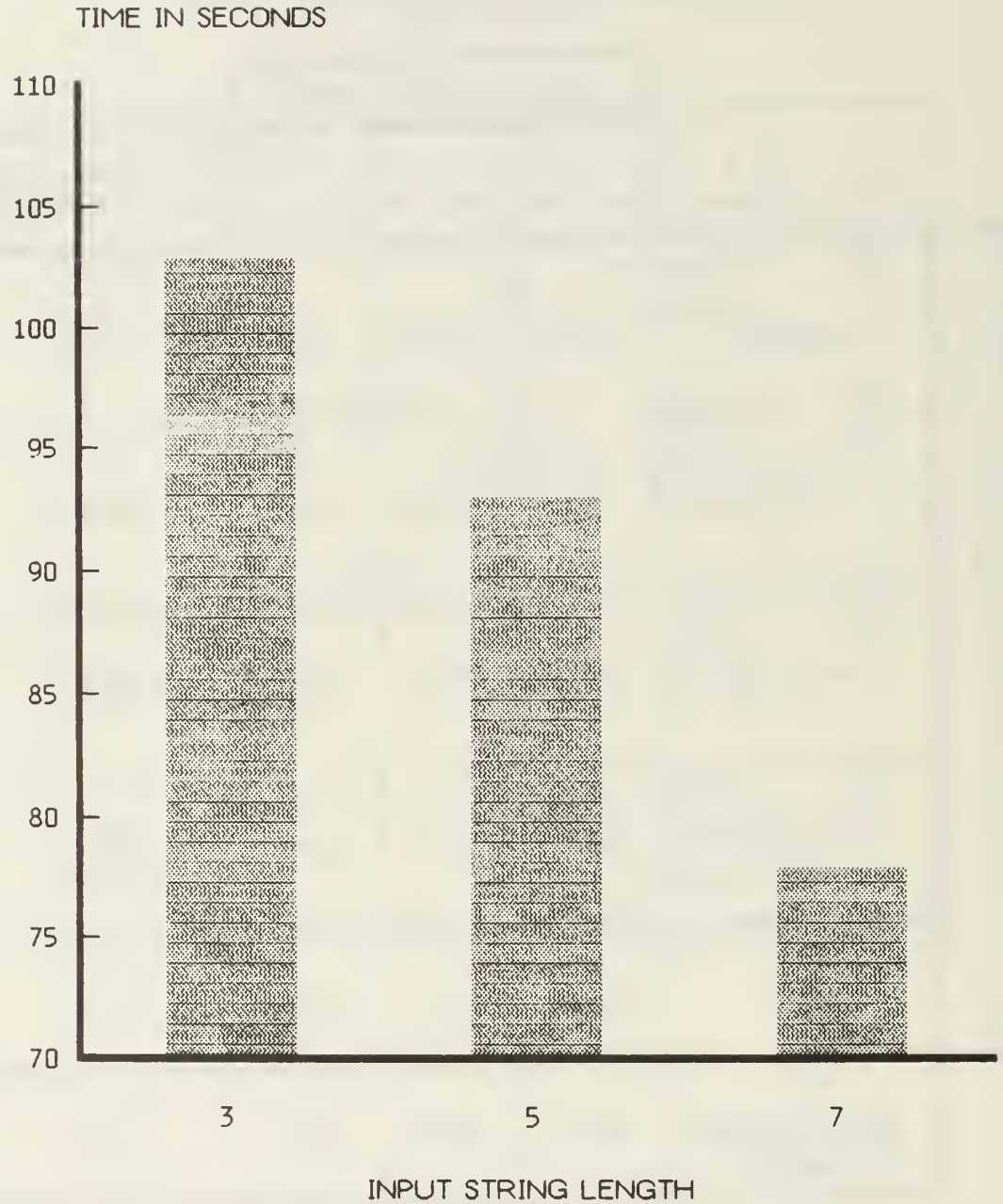

FIGURE 3-1.
MEAN TIME (IN SECONDS) BY INPUT STRING LENGTH

3.3 Total Errors

Table 3-3 presents the analysis of variance for total errors. A significant main effect of input string length was found ($F = 6.446$, $p < .01$). No other main effects or interactions were significant. Mean total errors for input string length by correction method are shown in Table 3-4. The main effect of input string length is shown graphically in Figure 3-2. Scheffe tests were performed to detect simple effects between input string lengths. Inputting digits seven at a time resulted in significantly fewer errors than when digits were input 3 at a time or 5 at a time ($p < .05$). However, the difference in errors resulting from inputting 3 digits at a time versus 5 digits at a time was statistically non-significant ($P > .25$).

Table 3-5 presents the results of subjects' choice of correction methods in the ALL condition and responses to the post-test questions. Subjects reported favoring CHANGE the most and BACKUP the least. However, in the correction condition in which all correction methods were available, CHANGE was used most and RUBOUT was used least.

TABLE 3-3

ANALYSIS OF VARIANCE SUMMARY TABLE
BY TOTAL ERRORS

SOURCE	df	MS	F
GENDER (G)	1	.12552	.474
ERROR	10	.26469	
CORRECTION METHOD (C)	4	.00389	.093
C G	4	.05264	1.264
ERROR	40	.04165	
INPUT STRING LENGTH (L)	2	.15870	6.446 *
L G	2	.00527	.232
ERROR	20	.06462	
L C	8	.02829	.849
L C G	8	.01521	.457
ERROR	80	.03331	

* P< .01

TABLE 3-4

MEAN TOTAL ERRORS (IN PERCENT) FOR
INPUT STRING LENGTH BY CORRECTION METHOD

		INPUT STRING LENGHT			
		3	5	7	$\overline{X}$ CORRECTION METHOD
CORRECTION METHOD	RUBOUT	4.545	3.979	2.315	3.613
	POSITION	4.307	3.837	3.523	3.889
	BACK-UP	4.121	6.839	4.262	5.074
	CHANGE	4.101	6.662	2.559	4.441
	ALL	4.631	5.394	3.772	4.599
	$\overline{X}$ INPUT STRING LENGTH	4.341	5.342	3.286	4.323 GRAND $\overline{X}$

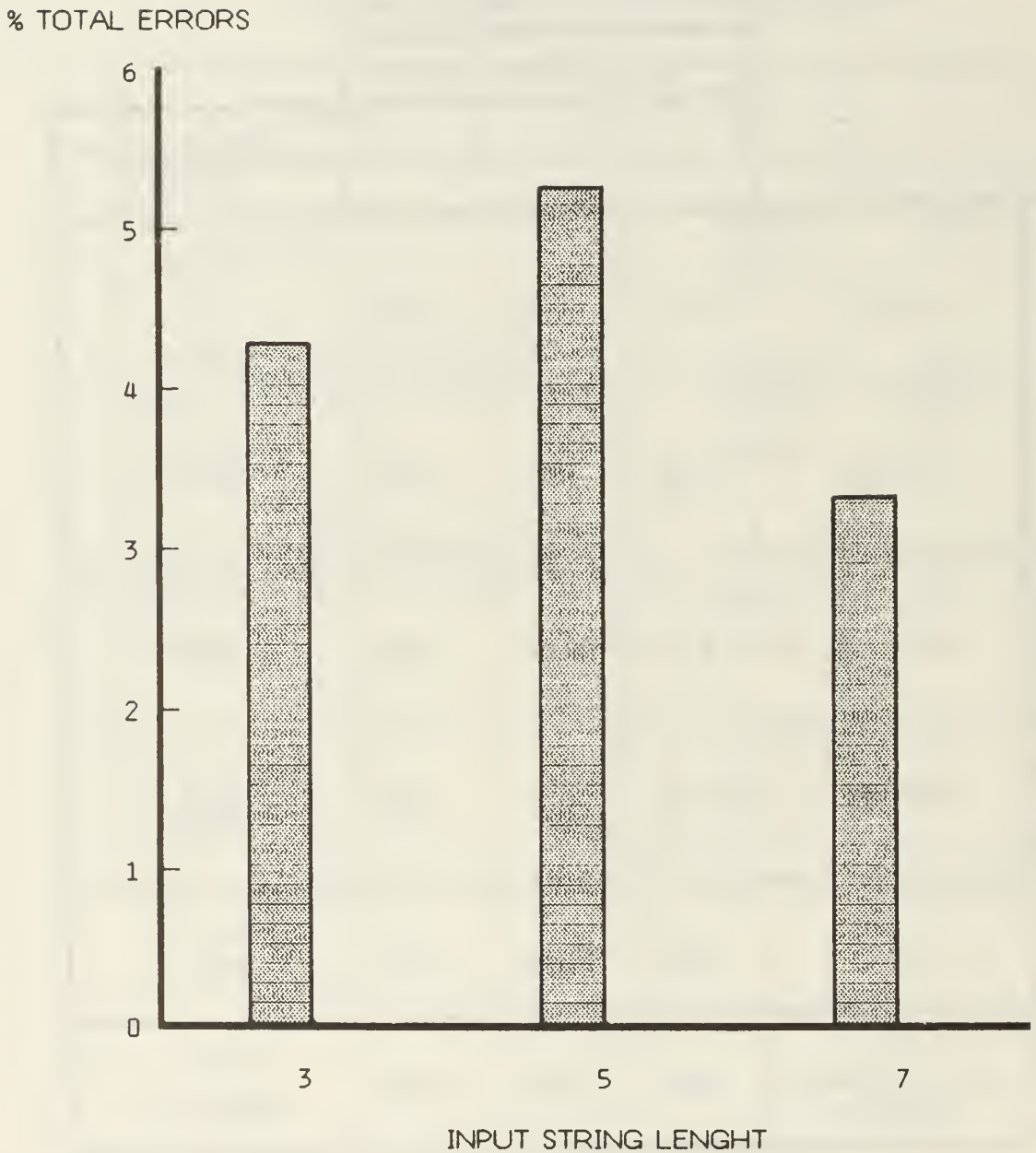

FIGURE 3-2.
MEAN TOTAL ERRORS (IN PERCENT) BY INPUT STRING LENGTH

TABLE 3-5

CHOICE OF CORRECTION METHODS AND
RESPONSE TO POST-TEST QUESTIONS

	CORRECTION METHOD			
	RUBOUT	POSITION	BACKUP	CHANGE
% THIS METHOD USED IN ALL CONDITION	12.7	22.7	14.6	50.0
% SUBJECTS WHO FAVORED THIS METHOD MOST	36	21	4	39
% SUBJECTS WHO WOULD OMIT THIS METHOD*	0	8	42	17

* 33% WOULD NOT OMIT ANY METHOD

4. DISCUSSION

This section will discuss the current findings with regard to the objectives put forth earlier in this report.

4.1 Effects of Input String Length

In terms of both accuracy and efficiency the results clearly demonstrated the advantages of inputting digits seven at a time compared to three or five at a time. This superior efficiency is probably a function of the relatively low number of both interstimulus pauses and errors, associated with the longer input string length. With an inter-stimulus pause of .8 seconds, error free pause times using the input string lengths of three, five, and seven, were 28 seconds, 16.8 seconds, and 12 seconds, respectively. Consideration of the inter stimulus pause reveals some noteworthy facts. If these pause times are deducted from the respective condition means, the differences among the resulting times are substantially reduced and in the case of input lengths three versus five, the direction of the difference is reversed (see Figure 4-1). However, the input string length of seven remains the fastest even if inter-stimulus pauses are eliminated completely, therefore, time is not solely a function of number of interstimulus pauses. Rather, time is a function of both number of interstimulus pauses and number of errors to correct.

Revised efficiency (time to input 105 digits minus error free interstimulus pause time) appears to be primarily a function of error rate. This supposition is supported by the data presented in Figure 4-2 which relates revised efficiency to error rate, by input string length.

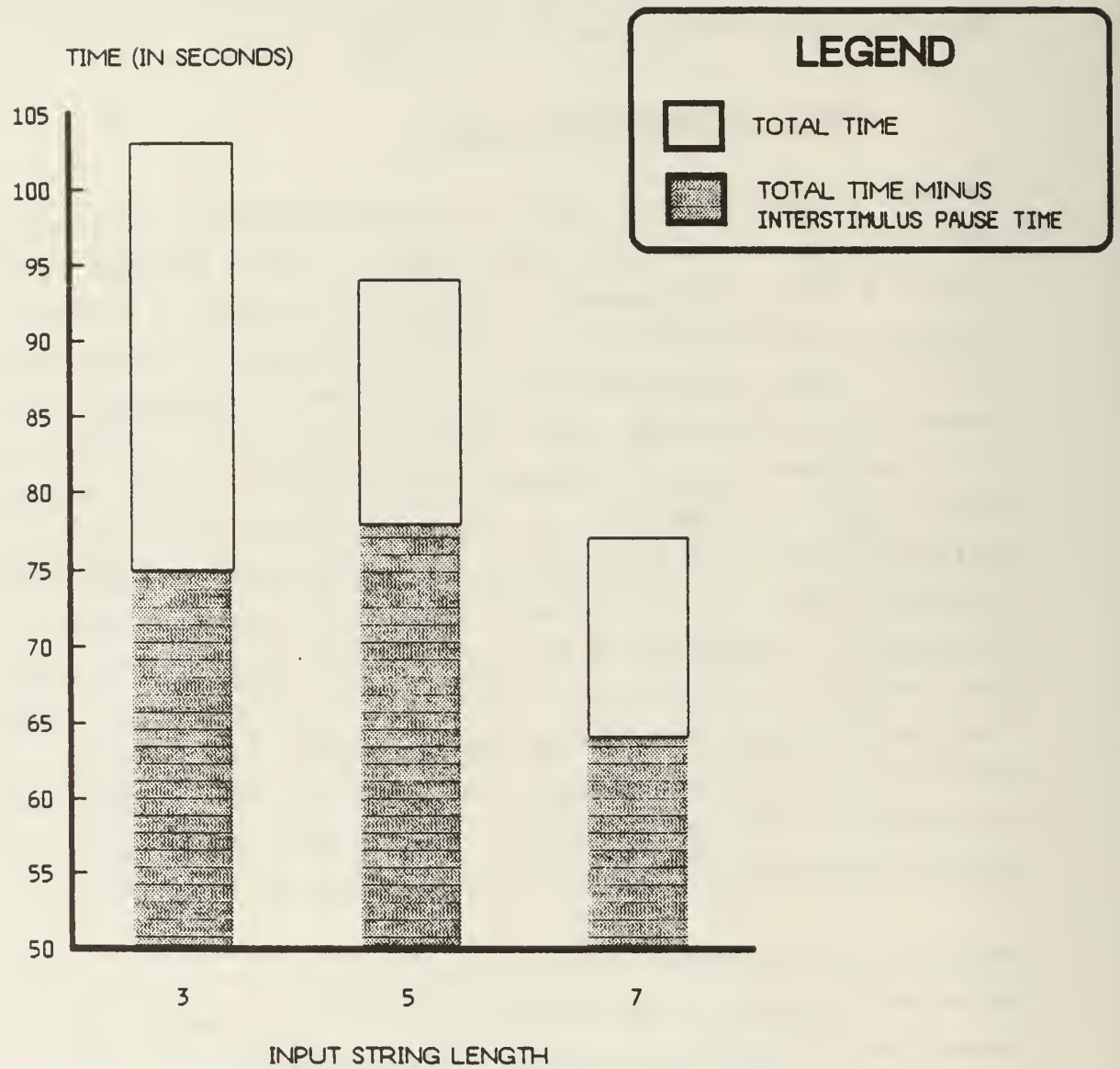

FIGURE 4-1.
TOTAL TIME AND TOTAL TIME MINUS INTERSTIMULUS
PAUSE TIME BY INPUT STRING LENGTH

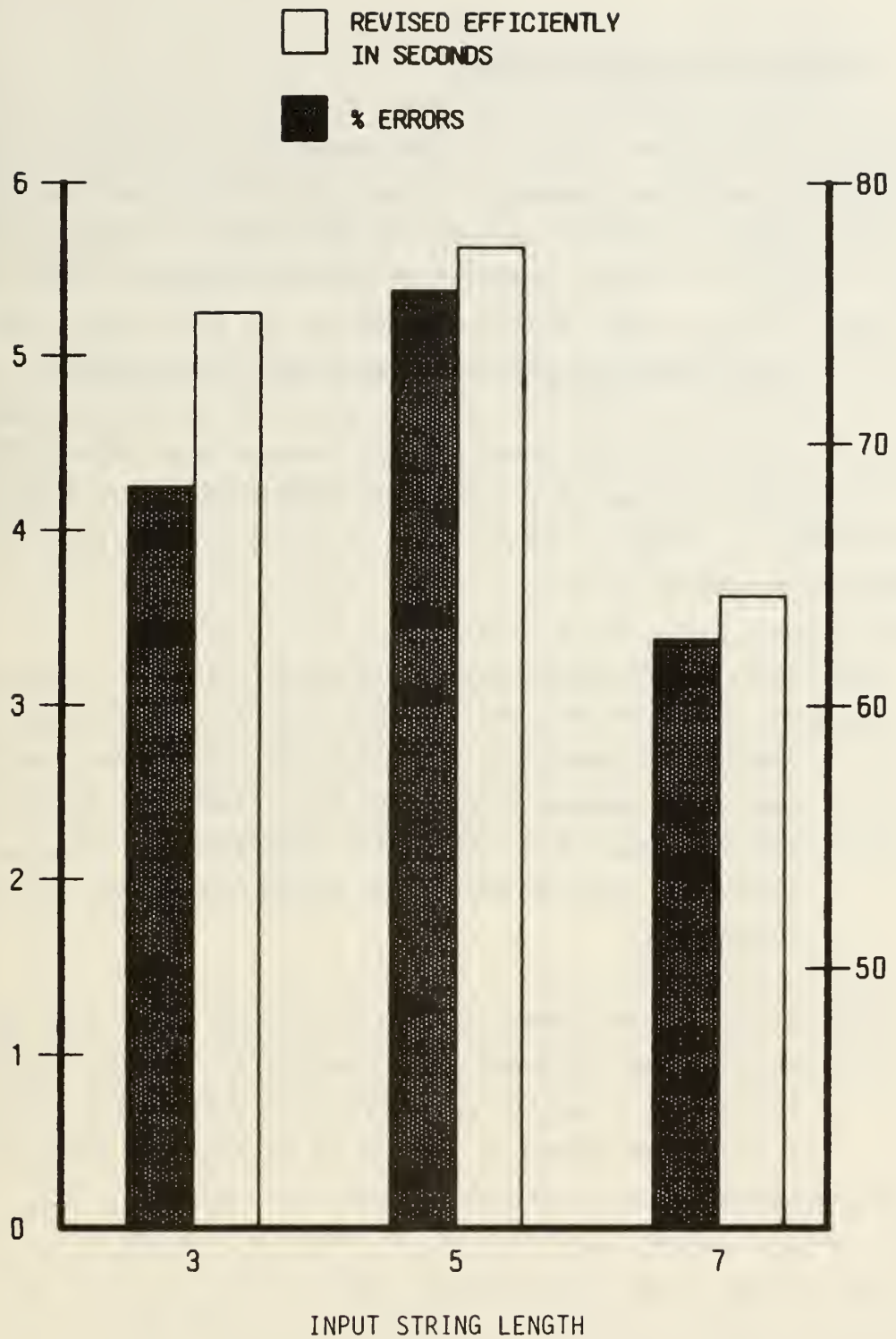

FIGURE 4-2.
REVISED EFFICIENCY AND ERROR BY INPUT STRING LENGTH

4.2 Effects of Correction Methods

There were no significant differences in accuracy or efficiency as a result of the various correction methods, and correction method did not interact with string length or gender. It was the experimenter's observation that the vast majority of errors consisted of one misrecognized digit in the spoken input string. Based on this observation one might expect the RUBOUT method to reduce efficiency, especially with the string length of seven, since the entire string had to be repeated after the correction command was spoken, and since this correction process involves two inter stimulus pauses compared to only one in all other correction methods. Hindsight and statements made by subjects provide some feasible explanations for the absence of this outcome:

(1) In using the RUBOUT method the subject did not have to search the output string for the specific error, determine its position or identity, and plug this information into the correction command format. This may have given RUBOUT a speed advantage over the other three methods which required the subject to perform additional mental processing and verbal formatting.

(2) A floor effect cannot be ruled out since there were very few errors under all conditions (grand $\bar{x}$ = 4.32% and RUBOUT $\bar{x}$ = 3.61%). As a result, subjects had an opportunity to implement a correction method an average of only four or five times per condition.

Subjects preferred the CHANGE method most and BACKUP the least and used the CHANGE method more than twice as often as any other method above given a choice of all four methods in the ALL conditions. Although statistically condition, which by far, the most subjects chose as the correction they would omit in a numerical data entry task.

4.3 Estimate of Recognition of a Continuous ASR Device

The recognition accuracy of the ASR system averaged 95.68%. In many ways this was a conservative estimate of the systems capabilities:

(1) While the system has the capability to adjust its gain level to speech versus background noise, this setting remains constant throughout training and testing. Therefore, sporadic noise changes such as those caused by the penumatic door opening and closing and the voices and typing of additional individuals entering (or leaving) the room are not accommodated by the gain level, and present a formidable challenge to the ASR device.

(2) Subjects were instructed to speak more rapidly than in normal conversational speed, increasing the degree of coarticulation and, presumably, making the task of speech processing more difficult than usual.

(3) Subjects spent only 20 minutes actually providing speech for template creation and, unlike many previous studies, "problem" words (words often confused with other words) were not retrained to an improved recognition criterion (Poock & Martin, 1983; Poock, Schwalm, Martin, and Roland, 1982).

(4) One error was recorded for each nonrecognition or correction made. In some cases, one or two errors may require several corrections. For example, the input string "1, 2, 3, 4, 5, 6, 7" is soken but the "1" is not recognized, constituting the first error. The ASR device now has "2,3 4, 5, 6, 7" in its recognition buffer and erroneously takes the next input (RESTART on background noise) as the seventh digit, constituting a second error. The resulting output buffer is "2, 3, 4, 5, 6, 7, 0" and in three of the four correction methods the subject had to make seven corrections to correct the entire string (e.g., POSITION 1, MAKE-IT 1, POSITION 2 MAKE-IT 2, BACKUP 1 MAKE-IT 7, CHANGE 6 MAKE-IT 4, etc.). As a result, one error of omission and one error of insertion lead to seven corrections, and would be recorded as seven errors rather than as two.

Finally, one factor should be noted that may have worked in favor of the ASR system. The vocabulary size of only 20 utterances is relatively small and the branching complexity of the grammar structure was fairly simple (see Figure 4-3).

4.4 Effects of Gender

As expected, gender did not significantly effect either accuracy or efficiency. This supports the findings of Batchellor (1981).

```
.  digit  .  digit  .  digit  .  digit  .  digit  .  digit

                                  OR

RESTART

                                  OR

RUBOUT

                                  OR

POSITION  .place   MAKE-IT  .digit

                                  OR

BACKUP    .place   MAKE-IT  .digit

                                  OR

CHANGE    .digit  <.which one>   MAKE-IT   .digit

digit =            .place =           .whichone=
   ZERO               ONE                FIRST-ONE
   ONE                TWO                SECONDONE
   TWO                THREE              THIRD-ONE
   THREE              FOUR
   FOUR               FIVE
   FIVE               SIX
   SIX                SEVEN
   SEVEN
   EIGHT
   NINE
   OH
```

FIGURE 4-3.
GRAMMAR STRUCTURE OF NUMERICAL DATA ENTRY VOCABULARY
FOR SEVEN DIGIT STRING

5. CONCLUSION

This exploratory study provided interesting and useful findings. The spoken entry of seven digits at a time proved significantly more efficient than shorter strings of digits. This effect prevailed despite the greater speech processing required by the Verbex, and the added processing imposed on the subject in repeating, checking, and correcting the seven digit strings versus the shorter input strings of three and five. The reason for such an outcome is currently unknown. The investigators only speculation is that the longer string (with resulting -- fewer pauses) was less prone to errors caused by the peek noises of the loud pneumatic door. Future research is suggested to test this speculation by repeating the basic experiment in a consistent sound level environment. In the meantime, input strings of seven digits are recommended for numerical data entry because of their efficiency in terms of error rate and minimal inter-input pauses. If input string length does interact with peek background noise, the use of seven digit strings becomes even more attractive.

Since no effects were associated with correction method, we suggest including CHANGE, RUBOUT, and POSITION as options for numerical data correction. BACKUP is deleted on the basis of high subject disappproval and low use.

The average recognition accuracy rate of over 95% is conservative and promising. We believe the outlook for ASR, especially in numerical data entry, is greatly improved with the advent of reliable continuous speech recognition capabilities such as those now available in the model used in our tests.

6. REFERENCES

Bradley, J.V. Probability; Decision; Statistics. Prentice Hall, Englewood Cliffs, N.J., 1976.

Bruning, J.L. and Kintz, B.L. Computational Handbook of Statistics (2nd ed.), Glenview, Illinois: Scott, Foresman and Co., 1977.

Miller, G.A. The Magical Number Seven, plus or minus two: Some limits on our capacity for processing information. Psychological Review, 1956, 63, 81-97.

Neter, J. and Wasserman, W. Applied Linear Statistical Models, Homewood, Illinois: Richard D. Irwin, Inc., 1974.

Nye, J.M. Human Factors Analysis of Speech Recognition Systems, Speech Technology, Vol. 1, 2, April 1982.

Poock, G.K. Voice Recognition Information Sheet. Naval Postgraduate School, 1982.

Poock, G.K. and Martin, J.M. Voice Recognition Performance with Naive versus Practices Speakers. Naval Postgraduate School Technical Report NPS55-83-016, June 1983.

Poock, G.K., Schwalm, N.D., Martin, B.J, and Roland, E.F. Trying for Speaker Independence in the Use of Dependent Voice Recognition Equipment. Naval Postgraduate School Technical Report NPS55-82-032, December 1982.

APPENDIX A

NUMERICAL DATA ENTRY VOCABULARY

Appendix A
Numerical Data Entry Vocabulary

1. ZERO
2. ONE
3. TWO
4. THREE
5. FOUR
6. FIVE
7. SIX
8. SEVEN
9. EIGHT
10. NINE
11. OH
12. RUBOUT
13. RESTART
14. POSITION
15. BACKUP
16. CHANGE
17. FIRST-ONE
18. SECONDONE
19. THIRD-ONE
20. MAKE-IT

DISTRIBUTION LIST

	NO. OF COPIES
Library, Code 0142 Naval Postgraduate School Monterey, CA 93943-5100	4
Library, Code 55 Naval Postgraduate School Monterey, CA 93943-5100	1
Director of Research Administration Code 012A Naval Postgraduate School Monterey, CA 93943-5100	1
Center for Naval Analyses 2000 Beauregard Street Alexandria, VA 22311	1
Professor G. K. Poock Code 55PK Naval Postgraduate School Monterey, CA 93943-5100	10
Naval Ocean Systems Center Code 421 San Diego, CA 92152	2

DUDLEY KNOX LIBRARY
3 2768 00329404 2